


Nitrogen Critical Loads: Critical Reflections on Past Experiments, Ecological Endpoints, and Uncertainties

Dose-Response:
An International Journal
January-March 2022:1–10
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/115593258221075513
journals.sagepub.com/home/dos


William M. Briggs¹ and Jaap C. Hanekamp² 

Abstract

Nitrogen Critical Loads (NCL), as purported ecological dose-response outcomes for nitrogen deposition from anthropogenic sources, play a central role in environmental policies around the world. In the Netherlands, these NCL are used to assess, via calculations using the model AERIUS, to what extent NCL are exceeded for different habitats as a result of different sources such as industry, agriculture, traffic. NCL are, however, not well defined, and are subject to hitherto unrecognized forms of uncertainty. We will address this with reference to a number of key studies that forms the basis for several NCL. We will subsequently propose amendments that could be applicable to future nitrogen studies and their enhanced relevancy in decision making.

Keywords

critical loads, nitrogen, uncertainty

Introduction

A “critical load (CL)” is an official level of exposure to a substance above which environmental harm is said is likely occur. These loads are mostly presented as atmospheric deposition rates of kilograms per hectare per year.

Nitrogen critical loads (NCL) have been at the forefront of governmental ecological protection policies in farming practices, industrial activities, urbanization, traffic, and so on. Their history dates back to at least the 1980s, when the first tentative experiments and observational studies were published on the effects of multiple air pollutants on the natural environment.

In this contribution, we will revisit a sampling of studies relied upon to set NCLs applied to different ecological endpoints as they are applied mainly in the Netherlands and Europe. We will analyze these studies from an informational standpoint; that is, we shall scrutinize the endpoints, the experimental set-ups, and inherent uncertainties either reported and applied or not. Before we can delve into the material, however, we first need to define what NCL are.

The following definition for NCL is from¹ “The term ‘critical load for nitrogen deposition’ means...: the limit above which there is a risk that the quality of the habitat will significantly be affected by the acidifying and/or eutrophication

influence of atmospheric nitrogen deposition.” This is somewhat loose, as the individual terms are undefined, leaving much room for differing interpretations.

De Vries et al.,² in their Critical Loads and Dynamic Risk Assessments Nitrogen, Acidity and Metals in Terrestrial and Aquatic Ecosystems, define CL, amongst which they include NCL, as follows: “Following the definition of a critical load by Nilsson and Grennfelt... the sustainability of the structure and function of an ecosystem is protected when a critical load is not exceeded by (atmospheric) deposition of pollutants, thus avoiding adverse effects and possibly irreversible damage in the future.” The extent of “adverse effect” is left undefined.

The study referred to in this quote is the 1988-report CL for Sulphur and Nitrogen edited by Nilsson and Grennfelt.³ This

¹Independent Researcher, Detroit, MA, USA

²University College Roosevelt, Middelburg, Netherlands

Received 29 November 2021; received revised 4 January 2022; accepted 4 January 2022

Corresponding Author:

Jaap C. Hanekamp, Environmental Health Sciences, University of Massachusetts, Amherst, MA, USA.

Email: j.hanekamp@ucr.nl and hjaap@xs4all.nl



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE

and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

report proposed the following CL-definition: “A quantitative estimate of an exposure to one or more pollutants below which significant harmful effects on specified sensitive elements of the environment do not occur according to present knowledge.”

The CL-definition as found in Nilsson and Grennfelt³ is still referred to as more or less authoritative. Be that as it may, these definitions are rather general in scope and need at least some specification with respect to endpoints and ecologies (types of ecosystems). The former are mostly described as eutrophication, acidification, and pressures on biodiversity (species richness). The latter refers to marine habitats (EUNIS class A), coastal habitats (EUNIS class B), inland surface water habitats (EUNIS class C), and so on. EUNIS stands for European Nature Information System.¹

Importantly, the endpoints of eutrophication, acidification, and pressures on biodiversity (species richness) interact with each other, according to the literature, in multiple loops and feedbacks.

In order to properly understand NCL as *actual measureable and decisional numbers of kg/ha/yr*, we propose a reverse-engineering strategy. That means that we take the actual NCL stated or used for some ecosystem and backtrack its origin from the available scientific literature.

Intriguingly enough, in the Netherlands NCL are reported as “unique singular values,” used ostensibly to determine deterioration of ecosystems once N-deposition, calculations made by AERIUS, exceeds these singular values. These unique singular values are portrayed as the best-estimate of the scientific state-of-art, for example, in¹ p. 13. About the on-line calculation tool AERIUS, see Ref.⁴ In one of the AERIUS documents (AERIUS, the calculation tool of the Dutch Integrated Approach to Nitrogen), N-deposition and NCL are discussed as follows: “AERIUS adds together the calculated amount of project-related deposition and the background deposition, and subsequently shows, per location, the total deposition in relation to the critical load. For ecologists, this is crucial information to assess the situation.”

In the present work, we address uncertainties and over-certainties specifically, as these concepts seem undervalued in the ecological studies we have scrutinized.

We examine what biological measures several authors have chosen below in their effort to understand the uncertainty in the measures themselves, and how these effect an understanding of the concept of critical load as such. The definitions of critical load given here are broad and do not imply any singular or specific or even range of policies that could or should be taken based on any critical load.

Nitrogen Critical Load Modeling

An approach to modeling nitrogen critical loads was given in,⁵ which we reviewed in.⁶ Briefly, in⁵ data was gathered either from papers detailing planned small-scale (time and space) nitrogen-added experiments or in papers detailing large-scale

observations. The data from each paper was comprised of a mix of plant growth, plant chemistry and nitrogen-uptake measurements. There was no strict consistency of these measures across the papers. Nitrogen deposition was either scaled up from the small time and space plot experiments or taken directly from the large-scale observations at the $\text{kg ha}^{-1} \text{yr}^{-1}$ level for all of the groups.

In most papers, controls, that is, no added- or low-N groups, were compared statistically with 1 of the several different plant growth or nitrogen-uptake measurements in N-added experiments, or in areas with lower (control) and higher atmospheric N. If the difference in these control-to-added-group measures gave a wee *P*-value, the difference was labeled (in effect) “harmful,” else it was labeled “not harmful.” The levels of nitrogen at the differences, or at the background if there was no wee *p*-value, was also noted. Finally, a logistic regression was created using these nitrogen levels and the labels “harmful” or “not harmful.”

If in the model an input level of nitrogen (in $\text{kg ha}^{-1} \text{yr}^{-1}$) gave 20% or more chance for “harmful,” that level of nitrogen was called “critical.” The *harmful*, again, was not consistent and was based upon a mixture of different outcomes, that is, various incommensurable measures of plant growth or chemistry.

Briggs and Hanekamp⁶ provide an in-depth statistical critique of this method, giving several suggestions for improvement in the modeling aspects alone. However, there are further troubling aspects about this approach than just the statistical modeling technical details as such.

Indeed, this approach, which was the first attempt to define uncertainties as such statistically, has a number of other obvious shortcomings. First, there was the mixing of planned nitrogen-added experiments with large-scale observations. The small plots (in time and space) were scaled up, but with no accounting for the added uncertainty inherent in this upscaling (e.g.),⁷ which must be substantial when moving from plots measured in the small square meters and short time periods to hectares over periods of one or more years. About this scaling, see more below. The experiments were controlled to some extent, but the large-scale observational measurements (typically country- or region-sized) were not (e.g.).⁸

The background levels of nitrogen deposition and its variability and seasonality were not well measured and reported across these papers. Some papers counted only wet deposition, some only dry, and some had the total. The uncertainty in the background levels was rarely given, or given only crudely (e.g.).⁹ Single, seemingly certain values of nitrogen were used for entire regions, with no real idea of changes in these numbers due to seasonality or geography (e.g.).¹⁰ As we also see and discuss below, the biological measures differed greatly as well. This is not inappropriate if decisions have to be made for each of these differing measures, but there is no justification given for combining different measurement types when deciding NCL.

Modeling is certainly a reasonable approach to quantify uncertainty in well-defined critical loads, while using homogeneous data. Decisions based on model probabilities must fully account for the costs and gains of those decisions, too, of course. It is unlikely a one-size-fits-all value of 20% chance for “harm” (defined above) to occur would be justifiable in every situation.

Here we do not attempt to critique any statistical model, or even any statistical testing techniques used within individual studies. Our concern comes *before* these modeling steps, however necessary they might eventually be. We want to build a consensus on the exact kinds of data needed, on the quantifiable and replicable definitions of harm, of critical loads, and all the elements that go into models.

Uncertainties

Our guiding discussion in this contribution is of Chapter 6 of the major report of Bobbink and Hettelingh,¹¹ “Effects of nitrogen deposition on mire, bog and fen habitats (EUNIS class D),” which “includes a wide range of wetland systems that have their water table at or above soil or sediment level for at least half of the year, dominated by either herbaceous or ericoid vegetation.”

Here we step through the most and paradigmatic important papers cited and highlighted in,¹¹ discussing them with respect to endpoints and uncertainties of the types mentioned above. Our interest is not in the specifics of the experiments outlined per se, nor on the precise details of the plant biologies. Moreover, we do not profess to have completed a universal survey of this large topic. But we do believe we have attended to the major components of uncertainty, all of which are imperative to recognize *before* any modeling can commence. Later we discuss what steps we believe need be taken subsequently.

Bragazza et al

One of the most informative papers addressing the various aspects of uncertainty is,¹² a work that examined ombrotrophic *Sphagnum* plants in 15 mires across Europe. Accordingly, we pay this paper the most attention.

Sampling was performed per country in each of three to six mires and three to six hummocks, all with dense cover of *Sphagnum* and low cover of vascular plants. Various *Sphagnum* species dominated particular mires. Each testing site was a 10 × 10 cm plot. It is unclear how long these plots were exposed, but the authors say sampling was done “mostly between mid-September and mid-October.” “Mostly” was not quantified.

Mean annual atmospheric measurements for precipitation, temperature, and N- and K-deposition were taken from third-party sources. One-number averages were used for entire regions and across whole years. For instance, Finland was summarized with having a background level of 0.2 g m⁻² yr⁻¹.

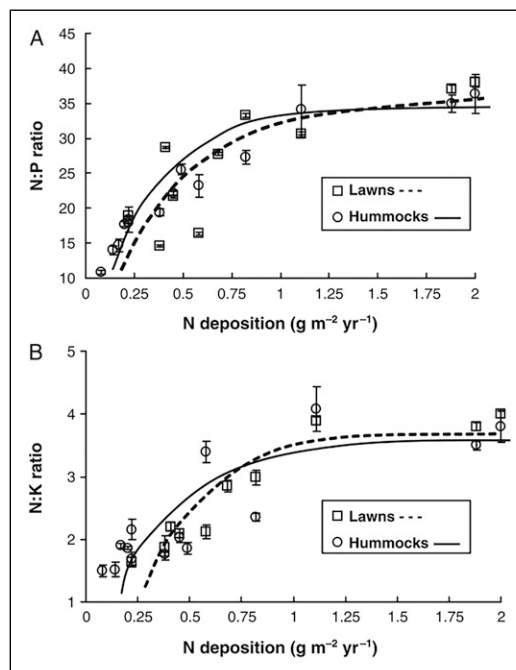


Figure 1. “Mean values (\pm SE) of (A) N:P and (B) N:K ratios in hummock and lawn *Sphagnum* plants at each mire in relation to atmospheric N deposition. Dashed and continuous lines represent the theoretical patterns based on regression model” from.¹² This figure is also cited in.¹¹

Sphagnum from these small plots were then analyzed, mainly for their N to P or N to K content ratios. These measured values did include standard error measurements, giving some indication of uncertainty of measurements over the 10 × 10 cm plots.

These chemical and biological measures were then input into regressions as outcomes with predictors being the country- or region-level single atmospheric N levels. It appears the locations or groupings of the individual plots were not controlled for, nor was species considered, though the difference between lawns and hummocks was noted in separate regressions. Not controlling for these various things would tend to, optically only, decrease any uncertainty in the resultant regressions, though we cannot here tell by how much, because no such uncertainties of the regressions were shown. That is, the model would not account for all relevant uncertainty.

The authors defined CL “as the amount of N bulk deposition above which *Sphagnum* plants experience nutrient imbalance to such an extent to greatly decrease the absorption of exogenous N. In this sense, N deposition levels above the critical load cause a N saturation of the *Sphagnum* layer, that is, the removal of N limitation associated with a decrease of N retention capacity. The atmospheric N input shifting ombrotrophic *Sphagnum* plants from being N limited to being P + K colimited is *c.* 1 g/m² year⁻¹, a critical load value consistent with the value suggested by other authors.”

This is a curious definition. It is true that if the local ecosystem has abundant N, then it could be P- or K-limited, but reaching these limits does not by itself imply any critical level has been reached. For instance, there could just be too low levels of P or K. It also does not appear to be a problem that plants do not store N at the same rates when N is abundant because, of course, N is abundant. In other words, just why these levels are critical is ambiguous.

In their report,¹¹ reference the central figure from,¹² also included here in Figure 1. This shows the N:P and N:K ratios of plant matter, plotted against the single-value atmospheric N estimates. Regressions are over-plotted separately for lawns and hummocks. Both the K and P limiting values are clear enough when atmospheric N exceeds about $1 \text{ g m}^{-2} \text{ year}^{-1}$. Once N exceeds this amount, plants have all the internal N they need, but could possibly use more P and K if it were available.

This does not make that $1 \text{ g m}^{-2} \text{ year}^{-1}$ a “critical” value, though. For if the environment had more available P and K, then the plants could conceivably take up more, and those N:K and N:P ratios would be pushed up to higher levels of atmospheric N. If anything, all this kind of picture indicates *by itself* is that once atmospheric N goes beyond a certain point, the effects of N are limited. They saturate with N, and perhaps not store as much, but then they do not need to, because there is plentiful atmospheric N to make up for whatever capacity to store is lost.

Based on very small sample sizes (number of mires ≤ 5) capitulum biomass (g dm^{-2}) was slightly less for atmospheric N greater than $1 \text{ g m}^{-2} \text{ year}^{-1}$ in hummocks, but slightly greater for lawns. The same was reported for stem volumetric density (g dm^{-3}). The signal is thus mixed; plus, there is no reporting of biomass or density at other values of atmospheric N. Which was scarcely possible given the small sample. And no other accounting of local conditions were given, so any indication of cause (N is not the only potential cause of change of the measurements taken) must be considered incomplete at best.

Berendse et al

A second representative paper is,¹³ which describes experiments in which $25 \times 37.5 \text{ cm}$ in situ plots in Finland, Sweden, Switzerland, and the Netherlands, each supplied with either extra CO_2 piped through hoses, or extra NH_4NO_3 added to water and sprayed on the plots. This was said to be equivalent of adding $3 \text{ g m}^{-2} \text{ year}^{-1}$ at three of sites, except the Netherlands, where it was equivalent of $5 \text{ g m}^{-2} \text{ year}^{-1}$. As in the previous paper, single-number summaries were used for atmospheric N, ranging from $.4 \text{ g m}^{-2} \text{ year}^{-1}$ (a site in Finland) to $3.9 \text{ g m}^{-2} \text{ year}^{-1}$ (a site in the Netherlands). These were not actual measurement, either, but output from a model (RAINS 7.2; see the paper for a description).

At the end of the experiment, a knitting needle was laid on the plots and the number of species touching it were recorded. *Sphagnum* dry weights and lengths were also sampled. The

added CO_2 and NH_4NO_3 experiments were compared with backgrounds. We do not here consider the CO_2 experiments.

Sphagnum production was greater in Finland in the added-N experiment contrasted with ambient conditions, and was less in other three countries. There was considerable variability across the four countries. N concentration was greater in *Sphagnum* plants in the added-N experiments in all four countries. There were more vascular plants noted in the added-N experiments than in ambient conditions.

The above-ground biomass was greater in two countries (Finland and Netherlands) and less in two others (Sweden and Switzerland) in the added-N experiments. But there was less below-ground biomass in all but the Netherlands, which had more. There was more standing dead and litter biomass (more carbon sequestration) in all four countries in the added-N experiments.

None of the differences in these experiments between ambient conditions were especially large, the signals are decidedly mixed, and as always it becomes difficult to know how to extrapolate the very small-scale experiments to entire regions, or even countries. It has the same difficulty as the previous paper in which the actual ambient levels of atmospheric N were unknown, and only assumed.

There is no way to draw, from this study, any critical value of N such that sufficient (or specific) undesirable changes have taken place. A critical value affecting biology could be defined, of course. Say, the percent coverage of non-*Sphagnum* plants exceeding some pre-specified level. But this would have to come with a plus-or-minus when extrapolating the experimental values to regions or whole countries.

Tomassen et al

Another paper was,¹⁰ describing an experiment in a simulated environment. Several $24 \times 24 \times 32 \text{ cm}$ jars into which extra N was injected were used. Varying amounts of N were added, from 0 to $4 \text{ g m}^{-2} \text{ year}^{-1}$, with “negligible” background deposition level of N. This negligibility was likely true because the environment was wholly artificial, unlike in other experiments which dealt with atmospheric background levels.

Differences were found at the highest N for plant biomass in two species of *Sphagnum*. But given the wholly artificial and small-scale nature of the experiment, it seems difficult to conclude as the authors do that “High N deposition levels do indeed appear to be responsible for the observed rapid vegetation changes in (actual) ombrotrophic bogs.” This might even be true, but guidance must be given by the authors of *how* to conform the measures taken in artificial environments and apply them to actual environments, all while accounting for the relevant uncertainties. However, this was not done.

Gunnarsson and Rydin

A fourth representative paper is¹⁴ that describes an experiment in which supplementary N from 0 to $10 \text{ g m}^{-2} \text{ year}^{-1}$ was

applied in *Sphagnum* hummocks and lawn communities on 20 × 20 cm plots. Ambient N ranged from 0.42 to 0.72 g m⁻² year⁻¹. This was a rare paper that also included some indication of uncertainty of these background figures, for example, standard deviation of ± 0.12 g m⁻² year⁻¹ for the ambient levels.

The signal of the experiment was mixed: “Sphagnum showed an increased growth in length with the intermediate N treatment, but in the second and third seasons the control treatment had the highest growth in length.” This was discovered by the use of an ad hoc formula as a function of length, density, and area in which samples were taken. That is, the result of the output of this formula became the key observation of study.

The authors say that up to a point “further N addition reduces growth,” but this could also be because of P or K limiting.

The authors also hazard a guess at a critical load, saying it is 1 g m⁻² year⁻¹, because this in their experiment matched the “optimal” growth of *Sphagnum*. But this was not optimal always and not everywhere. Growth at 5 g m⁻² year⁻¹ was higher for both locations studied than 1 g m⁻² year⁻¹, which was similar to the growth at 10 g m⁻² year⁻¹ in these small plots.

Shoot formation rates were maximized at 0 added g m⁻² year⁻¹ or three added g m⁻² year⁻¹.

Again, all of these measures may be important, either taken singularly, or taken together in some formulaic way. But that they change under varying N is, by itself, not important. The amount of change has to be demonstrated to be important by other means or arguments.

Breuwer et al.

The authors in⁹ conducted a greenhouse experiment, exposing different *Sphagnum* species collected at northern and southern Swedish locations to either ambient N or ambient N plus 4 g m⁻² year⁻¹ extra N. Temperature was also varied. The results were mixed. Biomass production was calculated by a formula with various plant measures as input. These formulas, which many authors use, are not uniform, and have many variations. This is not to argue any are right or wrong, as any of them might be right for a particular decision. But results from disparate formulas cannot be added together and directly compared without consideration of the uncertainties involved.

Northern species (*S. fuscum*, *S. balticum*) height increment and production changed with the changing temperature more than did southern species (*S. magellanicum* and *S. cuspidatum*). The southern species height and production changed more with the changing N. Some of the experimental containers “suffered from severe fungal infection.”

The ambient levels of atmospheric N the authors noted from other publications, and were from 0.3 to 0.6 g m⁻² year⁻¹. The ambient levels of N in the greenhouse was not noted. Once again, the result is that change was noted with

varying levels of N, but that what constituted harm, or how to extrapolate the small-scaled studies, was not laid out.

Limpens et al

The authors in¹⁵ also ran an experiment in controlled glasshouse conditions, examining the growth of *Betula pubescens* and *Molinia caerulea* in mixed *Sphagnum* samples gathered in the Netherlands and given 40 kg ha⁻¹ year⁻¹. They also examined growth of seedlings or sprouts of the same species, subject to infusions of 0, 40, or 80 kg ha⁻¹ year⁻¹. There was no apparent ambient level of N in the glasshouse atmosphere, or none was listed. Plant production, as in other experiments, was a formula based on various plant measurements. The same comments to that formula apply as above.

As for results, the differences in shoot mass by N, for instance, depended on the species (six were assessed). The signals were mixed. For some species, shoot mass was highest at 40 kg ha⁻¹ year⁻¹, and others at 80 kg ha⁻¹ year⁻¹, and one (*Molinia*) at 0 kg ha⁻¹ year⁻¹. Presumably there must, therefore, have been ambient glasshouse N. N in the interstitial water changed through time in a non-linear manner.

Because most of the largest changes were in *Molinia*, most of the discussion revolved around it and not the other five species. This is not inappropriate if this is the key species in some decision process. But it is also important to note that the results differed widely by species. And, as remarked above, no indication of how to extrapolate the experiment, which was in an artificial setting, to large scales was given.

Wiedermann et al

The work by¹⁶ is the most important paper in our review of this section because it attempted to answer how the small-scale experiments might scale up to regional or country levels.

They divided Sweden into four unequally sized regions with (from north to south) increasing levels of ambient atmospheric N, from which three mires were picked with a total wet plus dry deposition of N at three, 11, and 16 kg ha⁻¹ year⁻¹, with no uncertainty or variation given in these numbers, though the authors admit the difficulty of measuring dry deposition.

A small-scale controlled field experiment was also conducted using 2 × 2 m plots from which 0.5 × 0.5 m samples were taken. The experiments added-N at 0 (control), 2, 15, and 30 kg ha⁻¹ year⁻¹. S was separately varied, though this is not of direct interest to us. The ambient N was 2 kg ha⁻¹ year⁻¹, with no uncertainty given.

In the field experiments, vascular plant coverage increased on average in the samples from about 20% coverage at 2 kg ha⁻¹ year⁻¹ added-N, to about twice that at 30 kg ha⁻¹ year⁻¹ added-N. Total *Sphagnum* correspondingly decreased from just under 100% to just over 20%. These obviously do not sum to 100%, indicating these are relative area measures, and not total plant cover.

Table 1. Adapted from.¹¹ The Reliability codes are ### reliable, # quite reliable, (#) expert judgment. The ranges are not to be taken as indication of uncertainty, but should be used for different subkinds of ecosystems, as indicated in their text.

Ecosystem Type	kg ha ⁻¹ yr ⁻¹	Reliability	Indication of Exceedance
Raised and blanket bogs	5–10	###	Increase vascular plants, decrease bryophytes, altered growth and species composition of mosses, increased N in peat and peat water
Poor fens	10–15	#	Increase sedges and vascular plants, negative effects on peat mosses
Rich fens	15–30	(#)	Increase tall graminoids, decrease diversity
Montane rich fens	15–25	(#)	Increase vascular plants, decrease bryophytes

Observations taken *in situ* at three Swedish locations were also different with respect to coverages at the sites sampled. At the northern site samples with 2 kg ha⁻¹ year⁻¹ (again, no plus-or-minus to these were given), total *Sphagnum* was again about 100%, dropping to about 70% at site samples with 12 kg ha⁻¹ year⁻¹. Vascular plant cover rose from about 10% to just under 40%.

The authors called the patterns of changing cover in both the experiment and the *in situ* samples “analogous patterns,” which they are. The directions of change are certainly suggestive, but obviously many causal items of importance to growth differed between the artificial experiments and actual samples. The patterns could just as equally well be described by land use changes as by N differences. We do not claim this is the case, only note that it could be so and that the experiments did not rule this out.

Discussion of Bobbink and Hettelingh

From all these, and several other papers, Bobbink and Hettelingh derived what they classed as NCL. The following, in Table 1, is an adaptation of their Table 1 to indicate NCL for EUNIS code D ecosystems and their reliability.

The Reliability codes are (we are quoting) ### reliable, # quite reliable, (#) expert judgment. The ranges are *not* to be taken as indications of uncertainty, but are used for different subkinds of ecosystems, as indicated in their text. For example, for raised and blanket bogs, the authors say “use towards high end of range at phosphorus limitation, and towards lower end if phosphorus is not limiting” or “use towards high end of range with high precipitation and towards low end of range with low precipitation.” These judgments rely, they say, on several of the papers reviewed above (and of course others).

There is no other uncertainty in the numbers except for their judgments about the reliability. This is susceptible to several criticisms, which we lay out next.

- (1) No clear indication on what critical means. The official definitions of what critical loads is are to some extent clear with respect to political goals, but not clear with respect to repeatable or consistent measures on plant measurements or chemistry, for instance. We can see from Table 1 the changes vary widely. This

isn't a problem if it is these, and only these, measures upon which crucial decisions will be made. Otherwise, the exact physical or biological state that defines *critical* should be known and agreed upon before a critical load in any substance can be discovered. Is it growth rates of particular species beyond or below a certain point? Is it the amount of accumulated detritus? Of all plants? Only some? Is it a specific soil level of N? A ratio of N to K or P in dried plant matter? All species? Only some? It is a mix of species in which a favored species is too high or too low?

There does not have to be one singular definition, because there is not just one decision, or cost or benefit, in relation to N (or to anything biologically important chemical or biological component).

Whatever critical load is, it cannot change from one thing to another; change in measure, that is. Change in itself is neither good nor bad. That the different studies that showed, for instance, how the length of a particular *Sphagnum* species is changed on average in some way is not, by itself, of interest. It must be specified why some level of growth, if only growth is considered, is good or bad in some decisional manner.

The change has to be important in some named, clear, and measurable way. Naturally, this definition of critical is expected to change in different regions and times of years, and it might even vary by circumstances.

- (2) Statistical confusion. Statistics can give meaningful information about a critical load once it is defined, especially in quantifying its variability, seasonality and so forth. It should assist in quantifying uncertainty in critical load exceedance measures, as in *Nitrogen Critical Load Modeling* above. But one has to be cautious in using it to define critical loads itself.

There may be some confusion with “statistical significance.” In many of the experiments, biological measures were compared between conditions with added-N (at many levels) and a control, with N usually at ambient levels. If the difference between control and the other levels was “statistically significant,” that is a null hypothesis significance test evinced a wee *P*-value. This, by itself, is evidence of very little. Even

tiny differences, which would make no change in any possible decision, can be “significant.” A *P*-value is only that chance that, if no differences existed, a test statistic would exceed some level in new experiments (it actually means even less than this, but this is close enough, see).¹⁷

In any case, it is clear that statistical “significance” *by itself* cannot be used to define what is or is not critical. Critical has to be defined outside of any statistical test, and the definition must rely on the biology, chemistry, and even politics which underlie the N system. Once “critical” has been objectively defined, statistics can, of course, be used predictively and in other ways in concert with observations. But testing can never be the basis of the definition.

- (3) Lack of realistic studies. Most of the papers relied on very small areas in which precise measures were taken, with the results of these implicitly extrapolated to entire regions, or even whole countries. Other experiments were wholly artificial, inside glass houses for instances. It’s not that these experiments cannot provide useful information for designing large-scale measurements or experiments, but it cannot be argued with force that what happens in a controlled 10×10 cm plot is an error-free proxy for a region or country.

Even if what happens on these small plots can with confidence be extrapolated to large areas, that extrapolation comes at a cost in certainty. It cannot be maintained that, say, the exact proportion of species composition found on a 10×10 cm plot with a rigorously controlled addition of N (with the other components less well specified, or even unspecified) will be duplicated at a regional or country level.

There are certain mathematical techniques that can provide for the uncertainty due to extrapolations, but we would recommend against using them, because they themselves constitute a model, and would be just as untested as using the naive extrapolations. Again, large-scale experimentation is called for.

- (4) Full accounting of N. Nitrogen in soil less often mentioned in most of the experiments, though there are exceptions, for example,¹⁸ Again, regional, local, seasonal and other time-varying changes are never accounted for. Sources are speculated upon, but few confirmatory measures are taken. This, as we suggest below, calls for greater experimentation and observation.
- (5) Expert judgment unscrutinized In,¹¹ critical loads for various environments are given. For instance, for poor fens, the critical load is estimated to be $10\text{--}20 \text{ kg ha}^{-1} \text{ year}^{-1}$. The measures relied upon in these estimates are said to be “Increase sedges and vascular plants, negative effects on peat mosses.” The estimate is said to be “quite reliable” because it depends on the statistical hypothesis testing, criticized above.

But in other instances, such as for rich fens, which have a critical loads of $15\text{--}35 \text{ kg ha}^{-1} \text{ year}^{-1}$, and which relies on “increase tall graminoids, decrease diversity,” the levels are said to be made based on expert judgment. These judgments also rely upon certain small-scale experiments, but which do not have the same number of significance testing.

Indeed,¹¹ rely on “expert judgement” often in recommending critical loads. There is an appearance of variability noted in the judgments, but this is only apparent. The range is only given to further distinguish sub-types of environments. For instance, “high latitude or nitrogen-limited systems” should use the low end of $15\text{--}35 \text{ kg ha}^{-1} \text{ year}^{-1}$. What *exactly* is a nitrogen-limited system is not full specified.

We want to be clear. We do *not* wish to give the impression these expert-guided limits or definitions are incorrect *per se*. They may in fact be exactly correct. What we want to impart is that, given the evidence and objections made thus far, the certainty in these judgments is too high, and their bases too ambiguous. There needs to be a way to verify their accuracy (discussed), especially if and when expensive decisions will be made relying on them.

More experimentation and observations are thus needed.

Toward a Better Understanding of Uncertainty

The obvious intent of the studies discussed above is to infer causal effects of N. This can be done, with the obvious simplifying assumptions, for the small-plot experiments, where there is some form of control, at least with respect to the causative agent namely N. Extrapolating what was learned from the small plots to vast regions is another thing at together, of course, especially when trying to calculate the uncertainty.

What’s really needed, as we discuss below, are large-scale experiments on the same space and time scales that are important to decision makers. A step in that direction has been provided by¹⁹ and.²⁰ Both papers are similar, with²⁰ going farther in an attempted correlative quantification of N and certain biological measurements. The criticism we make is that cause is inferred by these quantifications, but should not be.

It’s worth going through the report²⁰ in some detail, because it sets the worthy goal of discovering large-scale so-called dose-response functions of atmospheric nitrogen and certain plant observables, such as “species richness.”

This makes a good first start at such an effort, especially considering much previous work, as we saw above, was on small-scale (20×20 cm) highly controlled plots, which were then extrapolated to regions, and even entire countries.

However, there are some areas of potential confusion which are addressed here in the spirit of constructive debate as much more work is needed to understand the science of N related to defining NCL as decisionals requiring nationwide investments.

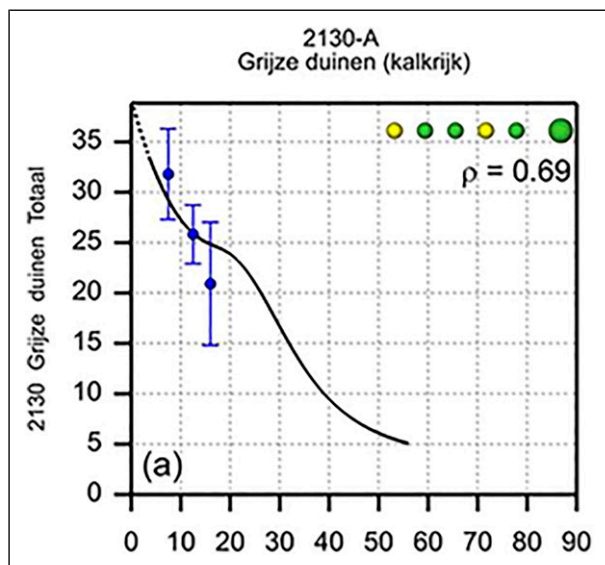


Figure 2. A so-called dose-response curve from.²⁰ The x-axis is in $\text{kg N ha}^{-1} \text{ year}^{-1}$. The y-axis is species richness, the model is in black, and the observations are in blue. The expert judgments are in green and yellow at the top. The correlation ρ measures agreement on model fit across experts.

Inappropriate Causal Inferences

The way both²⁰ and¹⁹ argue is the following. They observe various biological measures (such as “total species richness”) at several points in several different ecosystems. They also measure the corresponding N at those locations. They then build models, by ecosystem, that quantify things like species richness as a function of observed N.

They find, roughly, that species richness decreases with increasing N. Unfortunately, the temptation is to suppose that N is *causing* this decrease. This need not be so; indeed, it is likely not so in certain situations, which we discuss below.

Nitrogen is implied as causing lack of species richness; however, this might be an artefact of the way the data are analyzed. The data taken from the UK (from)¹⁹ is a good example of this. This shows for the ecosystem of dry dune grasslands, in total and separately for herbs and mosses, nitrogen deposition by “total species richness.”

We can here ignore the precise definition of “total species richness.” The suggestion that with and because of increasing N deposition “total species richness” decreases is there in this figure and in the text. But it is curious. The highest “total species richness” for both herbs and mosses is highest at, it appears, 5–10 $\text{kg N ha}^{-1} \text{ year}^{-1}$. The only areas shown in Figure 1 of Ref. ¹⁹, which shows a map of N deposition of the UK, where these levels of N are found are in far northern Scotland, and a thin band of coast in western Northern Ireland.

These coastlines are, of course, where dunes are found, so it is no surprise diversity of dune-based vegetation would be highest nearest the coast, and tapering off as the dunes transitions to other ecosystems.

Now the areas with least “total species richness” is for deposition rates of 15–17 $\text{kg N ha}^{-1} \text{ year}^{-1}$. These small areas border the coasts, reaching inland only a small way, such as middle northern Scotland, and on the inner western coast of Northern Ireland, a few scattered areas in northern Yorkshire, and a few very small areas along the coast of England. This is judging only by the picture: there may be other areas, but the point we make will be clear.

Obviously, as one moves inland, there will be fewer plant species that are associated with dunes, since dunes are found on coasts. It is therefore not necessarily N per se which is causing the reduction of diversity, but the lack of inland dunes causing declining dune-type vegetation. Certainly this alternate explanation is plausible, if not a better explanation.

Other figures repeat this “gradient studies” exercise for things like moss species and lichens in heather-dominated moors (Figure 5; p. 20)²⁰ and total vegetation and lichen in calcareous grasslands in the UK (Figure 7; p. 22).²⁰

In some of these plots, species diversity does not appear to be related to N because the standard deviation intervals overlap for most level of N deposition. In others it might, as in Figure 3. However, as Figure 1 of Ref. ¹⁹ illustrates, the same difficulty in assigning cause to N, and not to geography and land use, is found.

The conclusion that the “een negatieve correlatie gevonden tussen soortenrijkdom en/of samenstelling van de vegetatie en stikstofdepositie” (or “a negative correlation was found between species richness and/or composition of vegetation and nitrogen deposition”) is not wrong, but it does remind us that correlation is not causation, and should not be mistaken for it.

To turn this correlation into a causation, or at least turn it in that direction, more controlled data is needed. For example, nitrogen deposition rates at fixed locations in which the deposition rate is measured along species concentration in time. If the one vary together in time, then a case can be made for causation to be present. As it is, the signals so far seen could just be natural variation.

Inappropriate Dose-Response Inferences

The concept of “dose-response” is by definition causative. The data itself, mentioned in the previous subsection, is only correlative. Therefore, calling models of *in situ* observed N and things like species richness, relies on the implication that correlation is causation, that decreasing diversity is caused by increasing N.

Beyond all that, the so-called “response curves” themselves (p. 106) are of interest. See Figure 2.²⁰ The blue dots are the means of various biological measurements, with the standard deviations given, and the N (x-axis) divides the depositions into buckets in $\text{kg ha}^{-1} \text{ year}^{-1}$, with no uncertainty given. The dose-response curves are in black.

Instead of formally measuring curve fit (there are many methods), experts rated how well the curves represented the data (presumably by eye). These judgments are indicated in the colored dots at the top of the plots. Green is good, yellow so-so, red bad.

The authors did not appear to use any formal method of verification (the technical term for model diagnostics) because they had created the curves by maximizing a correlation coefficient. Perhaps they assumed formal methods of verification would thus be biased. This is somewhat, but not entirely true. Calibration, for instance, could still be checked. As could skill (measuring model performance against more simplistic, or even a constant dose-response model).

In any case, as is pictured here, the model curves extend far beyond the observed data. What is curious is that there is no way to say this is a good fit, or a bad fit. In fact, it is no fit at all. There are only three data points. And here the “dose-response” curve only comes close to 1 of these points. The curve is estimated for N levels out past 20, up to 60, but the data only goes to 15. But, even considering these criticisms, the curve was judged Good by expert opinion.

Looking closer, the curve is off by 5 units at the third and final data point, which does not sound like a lot until it's considered the data itself only spans 10 units. The curve is thus off by half here. The model cannot be considered good, but it might not appear that poor because the curve extends far beyond the observed N, stretching the curve.

Conclusion

The many small-scale studies done to identify critical loads of nitrogen are either over-certain, inconsistent, or not conclusive. Many are conducted on “plots” around the size of 20 × 20 cm, inside glass houses, with the results extrapolated and applied to large-scale regions, and even countries.

This obviously leads to great over-certainties. They are made worse because most, or even all, of these studies summarize ambient atmospheric nitrogen with one number, regardless of the size of the region of country. No account of variation, seasonality, or nitrogen sources is given.

The most difficult aspect is that these studies use inconsistent definitions of what “critical loads” are, defining them by crude statistical tests of the small-scale experiments. In effect, any change in these inconsistent outcomes over a “control” is taken to be “critical,” regardless whether the measures used (complex formulas of plant size, for example) would be interesting, actionable, or indeed cause any harm in actual locales.

There are two kinds of experiments that can and must be done to firm up all these uncertainties: (1) observational, to inform baselines and ranges, and (2) planned, to test causal ideas learned in the first step.

Observational

Nitrogen monitoring stations should be set up over a large area to measure wet and dry atmospheric deposition, and soil N content at several layers. The choice of location has to be for both adequate coverage and to inform decisions about N that must be made, as discussed shortly.

Those items thought or known to be associated with N must also be monitored. Obviously, N by itself is not usually of interest, but those things said to be causally affected by N.

Monitoring should be as finely grained as possible (large number of stations), with measurements taken (automatically) as frequently as possible. We need a clear indication of N flux, its variability, seasonality, and other characteristics.

The locations should be near or at those areas that are deemed crucial in some decisional sense, like agricultural fields, to measure, for example, crop yields; fens or bogs, to measure, for example, plant coverage or plant chemical content; lakes and ponds, to measure, for example, weed production. It must always be recalled that there are always costs and benefits from any policy. It should not only be costs driving decisions. The benefits of N should not be ignored, but usually are in efforts to paint N as a pollutant only.

Also, a change in any measurable item—from one location to the next, or one time to another, or because, say, farming has commenced—is not itself harmful or beneficial. Deciding harm or benefit is something that is “above” or independent of the measurements taken.

Certain atmospheric and soil variables should also be taken along with N. Temperature, precipitation and wind at a minimum, but also solar irradiance. Soil chemistry beyond N, such as potassium and phosphorus, and soil moisture can also be gauged.

Spatial time-series statistics techniques, such as Kriging, can be used to estimate N in those areas and times in which it was not monitored. These techniques can also be used to quantify uncertainty in the changes of biological measures (such as crop yield or Sphagnum coverage) with changes in N.

Planned Experiments

Standard techniques for medium-to large-scale field experiments can be done, using blocking to add N to co-located fields while keeping some as a control. The amounts added would be by wet deposition, that is, adding to water and spraying.

This would be useful for gauging distance and nearness effects of N to plots adjacent to experimental fields. The differences, if any, between biologically relevant measures in adjacent plots can also be ascertained. That is, the differences, if any, between those plots adjacent to where N was added, and those plots adjacent to control plots.

In this way the geographic extent of adding known amounts of N can be quantified.

Acknowledgments

Thanks to Geesje Rotgers for commenting on an earlier draft of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jaap Hanekamp  <https://orcid.org/0000-0002-6575-3658>

References

1. Van Dobben H, Bobbink R, Bal D, Van Hinsberg A. *Overzicht Van Kritische Depositiewaarden Voor Stikstof, Toegepast Op Habitattypen En Leefgebieden Van Natura*. Technical Report Alterra-rapport 2397, Wageningen. Wageningen, Netherlands. Wageningen Municipality; 2012.
2. Vries WD, Hettelingh J, Posch M. *Critical Loads and Dynamic Risk Assessments Nitrogen, Acidity and Metals in Terrestrial and Aquatic Ecosystems. Environmental Pollution, Volume 25*. Dordrecht, Netherlands: Springer; 2015.
3. Nilsson J, Grennfelt P. Critical Loads for Sulphur and Nitrogen. Report from a Workshop, Skokloster, Sweden, 19-24 mar 1988. Technical report. Copenhagen, Denmark: Nordisk Ministerraad; 1988.
4. Heer MD, Roozen F, Maas R. The integrated approach to nitrogen in the Netherlands: A preliminary review from a societal, scientific, juridical and practical perspective. *J Nat Conserv*. 2017;35:101-111.
5. Banin L, Bealey B, Smith R, Sutton M, Campbell C, Dise N. *Quantifying Uncertainty in Critical Loads*. Technical report, CEH Report to SEPA. Lancaster, UK: Centre for Ecology & Hydrology; 2014.
6. Briggs WM, and Hanekamp J. Outlining a new method to quantify uncertainty in nitrogen critical loads. *J. Philosophy* 2020. https://www.researchgate.net/publication/350313133_Outlining_A_New_Method_To_Quantify_Uncertainty_In_Nitrogen_Critical_Loads.
7. Aerts R, Wallen B, Malmer N. Growth-limiting nutrients in sphagnum-dominated bogs subject to low and high atmospheric nitrogen supply. *J Ecol*. 1992;80:11-39.
8. Wiedermann MM, Gunnarsson U, Ericson L, Nordin A. Eco-physiological adjustment of two sphagnum species in response to anthropogenic nitrogen deposition. *New Phytol*. 2009;181:208-217.
9. Breeuwer A, Heijmans MMPD, Gleichmanand M, Robroek BJM, Berendse F. Response of sphagnum species mixtures to increased temperature and nitrogen availability. *Plant Ecol*. 2009;204:97-111.
10. Tomassen HBM, Smolders AJP, Lamers LPM, Roelofs JGM. Stimulated growth of *betula pubescens* and *molinia caerulea* on ombrotrophic bogs: role of high levels of atmospheric nitrogen deposition. *J Ecol*. 2003;91:357-370.
11. Bobbink R, Hettelingh J-P. Review and revision of empirical critical loads and dose-response relationships. In: Proceedings of an Expert Workshop, Noordwijkerhout, Netherlands, 23-25 June 2010. Bilthoven, Netherlands: Coordination Centre for Effects, National Institute for Public Health and the Environment (RIVM); 2011.
12. Bragazza L, Tahvanainen T, Kutnar L, et al. Nutritional constraints in ombrotrophic sphagnum plants under increasing atmospheric nitrogen deposition in Europe. *New Phytol*. 2004;163: 609-616.
13. Berendse F, Van Breemen N, Rydin H, et al. Raised atmospheric CO₂ levels and increased N deposition cause shifts in plant species composition and production in Sphagnum bogs. *Global Change Biol*. 2001;7:591-598.
14. Gunnarsson U, Rydin H. Nitrogen fertilization reduces sphagnum production in bog communities. *New Phytol*. 2000;147:527-537.
15. Limpens J, Berendse F, Klees H. N deposition affects N availability in interstitial water, growth of sphagnum and invasion of vascular plants in bog vegetation. *New Phytol*. 2002; 157:339-347.
16. Wiedermann MM, Gunnarsson U, MatsNilsson ANB, Ericson L. Can small-scale experiments predict ecosystem responses? An example from peatlands. *Oikos*. 2008;118:449-456.
17. Briggs WM. *Uncertainty: The Soul of Probability, Modeling & Statistics*. New York, NY: Springer; 2016.
18. Berendse F, Lammerts EJ, Olff H. Soil organic matter accumulation and its implications for nitrogen mineralization and plant species composition during succession in coastal dune slacks. *Plant Ecol*. 1998;137:71-78.
19. Field CD, Dise NB, Payne RJJ, Btitton AJ, et al. The role of nitrogen deposition in widespread plant community change across semi-natural habitats. *Ecosystems* 2014;17(5):864-877.
20. Wamelink G, Goedhart P, Roelofs H, Bobbink R, Posch M, van Dobben H. *Relaties Tussen De Hoeveelheid Stikstofdepositie En De Kwaliteit Van Habitattypen*. Technical Report nr. 228471006. Wageningen, The Netherlands: Wageningen Environmental Research; 2021.